DOCUMENT RESUME

ED 268 183 TM 860 255

AUTHOR Wisniewski, Dennis R.
TITLE An Application of the Rasch Model to Computerized Adaptive Testing.
PUB DATE Apr 86
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Difficulty Level; Grade 5; Intermediate Grades; Item Analysis; *Latent Trait Theory; Mathematics Tests; Microcomputers; Multiple Regression Analysis; Reaction Time; *Response Style (Tests); *Student Attitudes; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Binary Search Scheme; Pontiac City School System MI; *Rasch Model; Student Needs Assessment Program Mathematics Test

ABSTRACT

Three questions concerning the Binary Search Method (BSM) of computerized adaptive testing were studied: (1) whether it provided a reliable and valid estimation of examinee ability; (2) its effect on examinee attitudes toward computerized adaptive testing and conventional paper-and-pencil testing; and (3) the relationship between item response time and item difficulty of the observations provided by the selected items. This research was conducted on a Commodore 64 microcomputer using the BASIC language. A paper-and-pencil test, the Pontiac Schools Student Needs Assessment Program-Mathematics Test, Grade 5 (SNAP5), was also administered to the 310 subjects, to obtain a valid estimate of examinee ability. This was a 44-item criterion referenced test. The BSM presented items until the last correct item was one less in order value than the lowest incorrect item. Results indicated similar BSM and SNAP5 scores for those students who took the BSM test first. Those taking the SNAP5 first had higher scores, a result attributed to the practice effect. Subsequent analyses indicated no significant differences between groups; correlations between BSM and SNAP5 were .69 and .68. There were no significant relationships between mathematics attitude or total response time and either ability measure. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED268183

An Application of the Rasch Model to Computerized Adaptive Testing

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D. R. Wisniewski
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

by

Dennis R. Wisniewski, Ph. D.

Pontiac City Schools

Pontiac, Michigan

TM 860 255

Paper presented at the Annual Meeting of the American Eduational Research Association, San Francisco, California, April, 1986.

AN APPLICATION OF THE RASCH MODEL TO COMPUTERIZED ADAPTIVE TESTING

Introduction

Perhaps no invention in history has made as rapid an impact on people's lives as the computer. According to one writer, if manned flight had developed as quickly as has computer technology, "an astronaut could have orbited the earth nine years after Orville Wright wobbled aloft at Kitty Hawk".

The application of computers in business, industry, and education continues to grow at an accelerated rate. Every person comes into contact with computers in some way. Because today's students live in a computerized world, they must learn to deal with these machines that have created the "Second Industrial Revolution".

Education has been affected by developments in computer technology. However, there has been a considerable difference between currently available computer technology and the era of technology employed in education. In short, there has been a reluctance to implement the computer into the school curricula, and changes in the implementation of technology in the educational arena have not yet arrived. This inertia will be removed if the cost of printed matter continues its exponential increase.

One area of education which has demonstrated a definitive need for computerization is testing. The computer can provide

the items to the examinee, analyze the responses, and score the test according to the student's individual ability. It provides this information virtually at the touch of a key, immediately upon completion of the testing situation. Such immediate feedback is vital to any form of instruction. Accordingly, several researchers have studied this and other related aspects (Reckase 1981, Weiss 1982, Green 1982). The computer is an ideal delivery vehicle, not only to manage the branching that is quite cumbersome in paper testing, but also to record the sequences of events that play an important role in scoring (Ward, 1985).

The argument in favor of individualized testing has been advanced by many testing and educational theorists. Notable among these is B. F. Skinner. Skinner (1968) stated, "Failure to provide for differences among students is perhaps the greatest single source of inefficiency in education" (p. 157). He further added, "The number of questions correctly answered on a test is quantitative...but it does not yield a measure appropriate to the behavior of the individual" (p. 172).

When a person has been placed into a conventional testing situation, he/she is expected to respond to all items. It is logical to assume that for nearly every examinee there are items to which he/she is virtually unable to respond. This is a function of the measurement variable which is constructed solely by the items in the test. It has been demonstrated by Birnbaum (1968) that the best information about ability estimates comes from those items which the examinee has probability of success equal to 0.5. Therefore, what function do these items at the extremes of the measurement variable have in assessing ability?

Item Response Theory and Adaptive Testing

There are a variety of item response theory (IRT) models which primarily differ in the number of parameters used to explain individual ability. An IRT model specifies a relationship between overall examinee test performance and a latent trait assumed to underlie the performance on the test. An IRT model is a mathematical model since the "unobservable" quantities are described by a mathematical function. The mathematical function is called an item characteristic curve. As a function it relates the probability of success on an item to the ability measured by the item set or the test it contains. It can be demonstrated that the shape of the item characteristic curve (ICC) will be invariant across randomly selected samples of examinees from a given population.

The one model which attempts to explain a response to an item by an individual in the simplest form and by the most elegant design is the Rasch model. It depends on only two parameters - person ability and item difficulty. The person's ability, the underlying latent trait, is set against the item's inherent difficulty. The functional value is the probability of success of answering the item correctly. This one-parameter model succinctly describes the latent behavior of an individual when he/she responds to an item in terms of probability, and treats the testing situation as a stochastic process. The items are assumed to be stochastically independent.

In an attempt to alleviate these problems, Rasch designed the concept of "specific objectivity". This concept summarizes the notion of stochastic independence and the notions of

instrument-free person measurement and person-free item calibration (Wright, 1977). The Rasch model has the following given assumptions about specific objectivity:

(1) item homogeneity, or unidimensionality,

(2) item characteristic curve,

(3) local stochastic independence.

If these assumptions hold, then the following elements are obtained:

(1) the raw score is a sufficient statistic for ability estimation,

(2) the comparison of two people is fully described by the difference in their abilities on the latent ability dimension,

(3) the estimation of the item difficulties is independent of the ability of the sample on which they were calibrated.

The Rasch model, the simplest logistic model, became the model of choice largely because of its inherent objective measurement characteristics. Reckase stated "the simple-logistic tailored testing procedure seems to be a reasonable alternative to the traditional classroom achievement test" (Rechase, 1977).

Clearly, the last two decades have been concerned with the notion of adaptive - tailored - testing. This potential for improved measurement was viewed with excitement, presenting a remarkably effective application of latent trait theory. It has been brought about through recent theoretical developments and through steady advances in computer technology. Tailored testing

is the fruition of the prediction implied by Green (1970) when he referred to: "the inevitable computer conquest of testing" (p.73).

As mentioned earlier, a major problem, arising because the same test items are given to all examinees, has been that many examinees receive items that are too easy or too hard. If a test item is too difficult, an examinee may resort to random guessing or omission of the response to the item. If the item is too easy, the information available in the response is no more functional that an omitted response. This results in scores containing more error at the extremes of the ability range, indicating that the test is most accurate for average ability examinees. This restricted range of accuracy is reflected in lowered test reliabilities. Thus, tailored testing procedures have been designed to minimize errors of measurement by administering to each examinee only items of appropriate difficulty. This is accomplished by selecting for administration items that maximize the information about an examinee's estimated ability level. The ensuing result of IRT-based adaptive testing is maximizing precision and minimizing time in the testing situation. As Millman as succinctly stated, "Item response theory-based systems of adaptive testing can substantially reduce average testing time without loss in precision, or can substantially increase precision for a fixed testing time" (Millman, 1984, p. 92).

An additional benefit of computerized adaptive testing is diagnostic evaluation of learning deficiencies. By reducing testing time and maintaining measurement precision, CAT can afford diagnosis of student achievement on the measurement

variable contingent on their responses to previous items.

The Study's Purpose and Limitations

The purpose of the research was to determine if the Binary Search Method (BSM) of computerized adaptive testing could accurately (i.e., reliably and validly) estimate an examinee's ability, to study the effects on attitudes of examinees towards computerized adaptive testing and conventional paper-pencil testing procedures, and to study the relationship between item response time and item difficulty of the observations provided by the various Binary Search Method selected items. The research was motivated in part by the problems related to the rapidly emerging computer technology, where development and implementation has outstripped research and evaluation.

While this study endeavored to make the BSM algorithm as valid and generalizable as possible, it nevertheless had certain limitations.

The use of the BASIC (Beginners All Purpose Symbolic Instructional Code) language has its limitations in its logical elegance and program structure. The use of a language such as PASCAL would have probably improved the sequencing speed and data storage of the program. Also, a machine language based algorithm would have provided considerable speed in presenting items to the examinee at the CRT, and it would have also provided much quicker access to and from the disk drive.

The Commodore 64 was limited in its computing power when compared to some of the larger microcomputer systems which had hard disk capabilities. The "64" was fairly suited to

individualized testing, but presented some problems when implemented on a large scale basis with many users accessing the system.

Description of the Study

This research study was experimental in nature. It was conducted to determine if a given method of testing, namely the Binary Search Method of computerized adaptive testing, could accurately estimate an examinee's latent trait on the single dimension of the test by sequencing the items within the neighborhood of the true latent trait parameter of the examinee. The single dimension, or measurement variable of the test, is defined by its items. The neighborhood of the latent trait parameter is the confidence interval estimate of the latent trait parameter as determined by the Rasch model estimation procedure (Wright, Mead, & Bell, 1979).

A paper-and-pencil test was utilized to obtain a valid estimate of the examinee's ability. The paper-and-pencil test used in this study was the Pontiac Schools Student Needs Assessment Program - Mathematics Test: Grade 5 (SNAP5). SNAP 5 was a criterion-referenced, locally developed test used primarily as an objective mastery assessment device in the spring with all fifth-grade students in Pontiac City Schools. Its items are representative of a domain of instructionally relevant tasks which form basic skills in fifth grade mathematics. This test was administered to all students who participated in the study. The SNAP5 test contained a total of forty-four items each with a different Rasch difficulty estimate. The range, in logits, of

the item difficulties was -1.8 to 1.9. This set of items was the "universe" of items for the experimental procedure.

The microcomputer system utilized in the study was the Commodore-64 with a CRT and a Commodore disk drive (model 1541). The software employed in the computerized adaptive testing on the microcomputer system was the Binary Search Method (BSM). This software, which selected and sequenced the items for CRT presentation, collected and recorded all pertinent data, and contained the test termination rules, was written by the author.

The BSM item selection and sequencing procedure is a simple but elegant binary search of the items, which have been ordered according to the items' interval-valued Rasch model estimates.

As with any binary search technique, the BSM procedure was based on the idea that from a given point of origin there were two ways to proceed pursuant to a reaction (i.e., response). This "binary" approach continues until a desired result has occurred. The BSM procedure begins the testing of an examinee at the "middle" of the ordered set of items. The use of a median difficulty item is a weak estimation point on the variable. However, it is the best possible one in the absence of prior information about the students' position on the variable being measured. That is, without some previously given ability estimate, a regression to the mean of the variable becomes the optimal place to begin item sequencing. The initial item was chosen from one of two items whose Rasch estimate was 0.00 or 0.03 on the variable measurement scale. These two initial items were of median difficulty, and were used assuming a normally

distributed population. The second item selected for presentation to the examinee was a function of the examinee's response to the initial item. If the response was correct, $X_{vi} = 2$, for person v and item i, then the next item was half way between the hardest item and the last item correctly answered. If the response to the second item presented was incorrect, $X_{vi} = 1$, then the third item for examinee response was halfway between this last incorrectly answered item and the last correctly answered item. This binary search for items continued until the last correct item is one less in order value than the lowest incorrect item. Accordingly, this establishes the termination criteria. Figure 1 lists two illustrations of the item sequencing.

It was clear that the optimal point to begin testing was with a median difficulty item, it also was inherently the weakest. To combat the inherent measurement error which could result, two special conditions, I and II, were built into the BSM algorithm. Condition I, the so called false-positive initial response condition, sequenced the examinee to the lower difficulty half of the variable in the instance where the initial item response was correct but was followed by three successive incorrect responses. Condition II, the false-negative initial response condition, performed an equal but opposite function in item sequencing.

Illustration 1. Regular Sequencing.

Attempted items[a]: 21, 12, 16, 14, 15

Response values[b]: 1, 2, 1, 2, 2

Graphics on item sequencing: top arrows for correct response, bottom for incorrect.

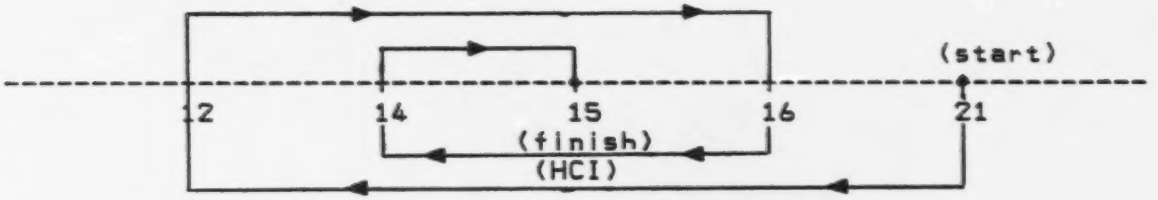
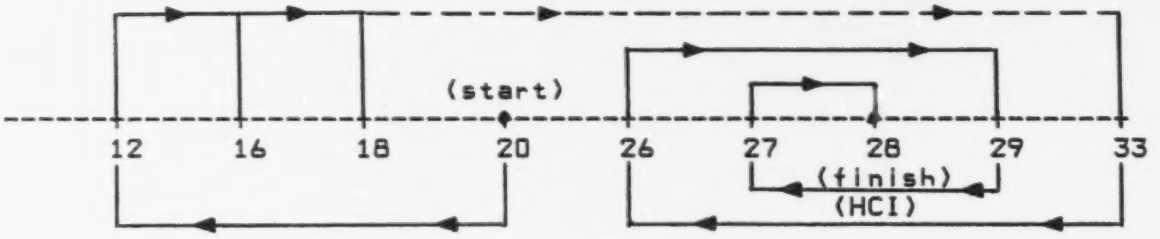


Illustration 2: Special Condition II Sequencing.

Attempted items: 20, 12, 16, 18, 33, 26, 29, 27, 28

Response values: 1, 2, 2, 2, 5[c], 6, 5, 6, 6

(this illustrates a false negative initial response)



Note: a - item presentation sequence to the student. The numbers are the BSM enumeration of the Rasch item difficulties in ascending order.

b - answered correctly if even value, incorrectly if odd

c - coding value for special condition sequence

Figure 1. Diagrams illustrating the BSM algorithm for normal and special condition item sequencing.

The subjects in the study were 310 fifth grade students all of who took both forms of the 44-item mathematics test. These examinees were randomly separated into two groups, A and B.

Group A consisted of 154 examinees, while group B and 156. Group A took the BSM procedure form of the test first, while Group B took the traditional paper-pencil form. Approximately five weeks later both groups took the alternate procedure. Group B had thus seen all 44 items on the initial testing. This practice effect demonstrated itself in the results.

The test's domain, which effectively defined the measurement variable, contained items which were part of the instructional objectives in mathematics to be covered during the fifth grade. Since both forms of the test were administered from mid-October to mid-November, both test forms served as a pre-test.

Results

A t-Test was conducted to determine if any statistically significant difference existed between the ability estimates when made by the BSM and SNAP5 procedures.

The null hypothesis was accepted for Group A, while being rejected for Group B. Table 1 displays the t-Test results for the two-tailed probability.

For Group A, the mean and standard deviation for the BSM procedure was 95.98 and 5.49, while the mean and standard deviation on the SNAP5 was 96.16 and 6.68. A two-tailed t-Test was not significant at $p<0.05$.

For Group B, the mean and standard deviation for the BSM procedure was 96.48 and 5.08, while the mean and standard deviation on the SNAP5 was 95.75 and 5.17. A two-tailed t-Test was statistically significant at $p<0.05$.

The findings indicate that the BSM procedure approximated the score on the paper-pencil test for those Group A examinees

who took the computer adaptive test procedure first. This group saw only five to nine of the total forty-four items which appeared on the paper-pencil test.

The significantly higher scores on the BSM procedure for Group B examinees, who took the forty-four item paper-pencil test first, were attributable to practice effects occurring between the two test procedures. The time interval for the two procedures for both groups was approximately five weeks.

Supplemental analyses indicate no significant differences between the two groups on their respective test procedures.

Table 1

Correlated t-Test for Each Group on the Two Different Procedures

		Ability Score			
Group	Procedure	Mean	S.D.	t Value	Significance
A	BSM	95.98	5.49	-0.44	n.s
	SNAP5	96.16	6.68		
	d.f. = 154				
B	BSM	96.48	5.08	2.22	*
	SNAP5	95.75	5.17		
	d.f. = 156				

*p<0.05, n.s. = not significant

Pearson product-moment correlation coefficients were calculated to determine if a relationship existed for either

group between the BSM and SNAP5 ability estimates.

The null hypothesis was rejected for both groups. The correlations were significant at the 0.01 level.

The correlations between the BSM and SNAP5 procedures results for both groups A and B were significant at $p<0.01$. An $r = 0.68$ was reported for Group A, while an $r = 0.69$ was reported for Group B.

Of importance, as seen in Table 2, was the maximum, minimum, and range of scores of the two groups on both test procedures. The initial-test procedure results for both groups were similar, although the BSM procedure estimates for Group A were consistently lower. These lower estimations were not statistically significant as demonstrated by other analyses found elsewhere in the findings. The range of scores and their proximities at the extremes suggest some potential relationship; no concommitant variable or variables were identified on which to match the groups' examinees. Thus, no test of significance was attempted. The scores are stated in transformed logit values of ability estimates.

Table 2

Maximum, Minimum, and Ranges of

BSM and SNAP5 Scores

Group	Procedure	Max.	Min.	Range
A	BSM	109.6	79.4	30.2
(N=155)	SNAP5	112.8	69.5	43.3
B	SNAP5	112.8	84.4	28.4
(N=157)	BSM	115.1	81.4	33.7

In order to determine if fifteen to twenty percent of the SNAP5 items, as selected by BSM, would render a reliable ability estimate, linear regression was employed using the BSM as a predictor of the SNAP5 criterion variable.

The null hypothesis was rejected. The findings indicate a reliable measure of estimated ability for examinees in Group A. The same reliability was indicated for Group B.

The prediction equations for Group A and Group B are listed below, respectively (see Equation 1 and Equation 2). Y' was the SNAP5 criterion variable and X was the BSM predictor variable.

$$Y' = 17.1 + .83X \quad (1)$$

$$Y' = 28.8 + .69X \quad (2)$$

Pearson product moment correlation coefficients were calculated for each group. A value of $r = 0.68$ was found for both groups, which was statistically significant at $p<0.01$.

In order to study specific changes in mathematics attitude due to exposure to computerized adaptive testing, the twenty

mathematics attitude items of the Mathematics Attitude Survey (MAS) were divided into the ten positive and ten negative response-type items on the survey. This was done to determine if any significant difference existed due to computerized adaptive testing on either positive-response or negative-response attitude types. A correlated t-test was employed to compare these various mathematics attitudes before and after exposure to computerized adaptive testing within each group.

The hypothesis was not rejected at the 0.05 level of significance for positive mathematics attitudes for either group, nor was it rejected for negative mathematics attitude for Group B at the same significance level. The hypothesis was rejected at the 0.01 level of significance for Group A negative mathematics attitude. A correlated t value of -0.06 was determined for positive mathematics attitude for Group A. A value of 0.41 was calculated using a correlated t-test for positive mathematics attitude for Group B. Neither value was statistically significant at $p<0.05$.

A correlated t value of 2.83 was calculated on negative mathematics attitude for Group A. This value was statistically significant at $p<0.01$. The similarly calculated t value of -0.58 for negative mathematics attitude was not significant at $p<0.05$.

The t value of 2.83 demonstrated a significant decline in negative mathematics attitude by examinees in Group A. It was implicit that exposure to computerized adaptive testing had a significant effect on negative mathematics attitude for those exposed to the BSM procedure first. The same is not true for

Group B examinees who were exposed to the SNAP5 procedure first.

A Pearson product moment correlation coefficient was calculated for each group to determine if any significAnt relationship could be found between total item response time and the examinee's BSM ability estimate. For Group A, a value of $r = 0.11$ with $r^2 = 0.013$ was not significant at $p<0.05$. Likewise, for Group B, $r = 0.03$ with $r = 0.03$ with $r^2 = 0.001$ was not significant at $p<0.05$.

These findings suggest that individual response time is probably a personal variable independent of the individual's latent ability on the test's dimension. Since the BSM ability estimate is identical to the item difficulty of the highest correctly answered item, it was inferred that the item difficulty and response time were not significantly related.

A multiple regression analysis was employed with BSM as the criterion variable and SNAP5 and mathematics grades as predictor variables. This analysis was made to see what relationships existed between the BSM estimate and the SNAP5 ability estimate and the examinee's previous mathematics grades.

The hypothesis was rejected for both groups. The multiple R values for both groups were 0.68, which were significant at $p<0.01$. The F ratios for both groups were also significant at $p<0.01$.

An examination of the findings shows a non-significant contribution to prediction of the BSM ability estimate made by the previous mathematics grades. The contribution made by the SNAP5 ability estimate was significant.

These findings coincide with the previous hypothesized

findings. Prediction of one testing procedure by another provides for a measure of linearity of the prediction equations. Further, the major contributions in predicting either score are by the other testing procedure score.

There were two post hoc analyses which related significantly to the postulated results of this study. The first result dealt with how well the highest correct item (HCI) of the BSM test procedure approximated the Rasch ability estimate as determined by the SNAP5 paper-pencil test procedure. The relative efficiency function, $I = 400P(1-P)$, where P is the probability of a correct response, was used to determine this HCI fit.

The probability of a correct response, P, is a function of the difference between the examinee ability estimate and the difficulty of the item encountered by the examinee. In determining the relative efficiency, I, the ability estimate was determined by using the SNAP5 raw score and the encountered item difficulty by the HCI of the BSM procedure. The SNAP5 ability was considered the standard ability estimate against which the BSM estimate was compared.

Table 3 displays results related to the first post hoc analysis. The results are separated by group because of the different test conditions for Groups A and B. The results presented in the table indicate that the BSM HCI did well in approximating the SNAP5 paper-pencil test score. For instance, an information value of 90% or more indicates an absolute difference between estimated ability and item difficulty of less than 0.7 logits. The table displays that over 87% of the

examinees from both groups were in this category. This information cutoff of 90% also indicates that it is very difficult to detect item misfit from the examinee's latent ability measure on this test's dimension.

Table 3

Frequency Distribution of
Relative Information Percentages
for Groups A and B

Percentage Range	Group A	Group B
0 - 9	0	0
10 - 19	1	0
20 - 29	0	0
30 - 39	1	0
40 - 49	0	0
50 - 59	1	1
60 - 69	1	1
70 - 79	6	8
80 - 89	9	8
90 - 99	136	139
	N = 155	N = 157

The second post hoc result was an examination of the relationship between the Rasch difficulty estimates used for the BSM procedure and those of the examinee item response vectors on SNAP5. The BICAL 3 item calibration program (Wright, Mead, Bell, 1979), was used to find item difficulty estimates on the results of the pencil-paper, SNAP5, test. Utilizing the 310 response vectors of the studys' subjects, a correlation of 0.92 was

determined. The value is quite significant considering the different testing times, and the different score distribution of the two calibration samples. Item misfit was not significant and as such did not contribute to any serious reduction in estimation precision.

Discussion

The BSM procedure showed a strong correlation with and good prediction of the SNAP5 procedure of ability estimate. Multiple regression analyses demonstrated repeatedly that the strongest contribution to prediction was by the BSM ability estimate. The prediction to the SNAP5 estimate by the BSM procedure lost some of its accuracy by the varying testing milieu which existed as an uncontrollable contaminant in the various classes. This milieu was largely, if not solely, determined by the classroom teacher. This was demonstrated by the consistency of ability estimate made by the BSM procedure, and the inconsistency of ability estimate made by the SNAP5 test. In those classes which showed a low correlation between the two procedure estimates, the testing milieu was markedly poor. The correlations between the two procedures for all twenty classes ranged from 0.23 to 0.87. From these statistics it appears that the teachers influence in the testing milieu had a significant influence on the performance of an examinee on a paper-pencil test. This assertion was evidenced by an informal survey of these teachers by the author. It was determined that for those teachers who established a good testing environment also demonstrated a high correlation between the results of the two procedures.

It was found that a proper, teacher-developed testing environment was not necessary when the BSM procedure was employed. The computerized testing procedure caused the necessary and sufficient testing environment for the examinee. It controlled the presentation rate of the items, focused the attention of the examinee to the task at hand, demonstrated its "patience" by allowing ample time for response and focused the examinee's attention to a single test item.

In the attitude survey information, there was no significant difference before, between and after either testing procedure. This relates well to the finding that there was little or no correlation between positive mathematics attitude and either of the ability estimating procedures.

The total amount of response time on the computerized procedure demonstrated no correlation between either of the ability estimating procedures. It was initially hypothesized that a strong negative correlation would be demonstrated, but this was not substantiated by the data. It would appear that the amount of time an examinee takes to respond to an item may be dependent on some unhypothesized variable or variables not accounted for in this study.

The finding that no significant relationship exists between total item response time and BSM ability estimate further underscores the importance that time not be used as a constraint when certain objective test measurements are made. The examinees perform nearly as well when provided with fewer test items and reasonably unlimited response time.

Whether or not time factors of individual response set are significantly related with conventional paper-pencil tests in general was not determined. The length of time spent by each examinee on the SNAP5 procedure was not one of the data points.

The results of this study showed a statistically significant relationship between the BSM and SNAP5 ability estimates. Although the reliability coefficient was not exceedingly high, there was support for the assumption that the BSM procedure produced a reasonably accurate estimate of the SNAP5 ability measure. This reasonably accurate estimate was made using only fifteen to twenty percent of the available item pool and was considered a valid finding.

Among the 310 examinees who were administered the BSM procedure, 34, or slightly less than eleven percent, responded to additional items beyond the normal five or six. This occurred because the BSM algorithm determined a Condition I or Condition II item sequencing for these examinees. As described above, Condition I was a false negative response on the initial response, while Condition II was a false positive one. The BSM and SNAP5 ability estimates for these particular examinees were compared separately for significant disparity between their scores. This comparison demonstrated virtually the same correlation between the ability estimates of this subgroup as the correlation between the estimates of the entire sample. Also, it was of interest that the percentage of false negatives and positives in a Monte Carlo simulation of the BSM procedure was virtually identical to that of the actual field study.

The validity of this study was supported in part by the post

hoc analysis of the relative efficiency of the HCI to estimate the SNAP5 estimate of ability. The large percentage of high efficiency demonstrates that the BSM algorithm sequenced the items in such a way as to arrive at an item whose difficulty was congruent to the latent ability measure of the examinee. One of the salient features of the Rasch model is that person ability measurement is best when item difficulty and person ability parameters are equal or nearly so. Concomitantly, as the relative efficiency approaches one hundred percent, the ability and difficulty parameters near equality.

Since an examinee's previous mathematics grade had little predictive value to either the SNAP5 or BSM ability estimates, its use as prior information for initializing the sequencing the items was seen as minimal. The fact that the BSM procedure significantly approximated the SNAP5 ability estimate without use of any prior information is of value to the conclusions of this study.

BIBLIOGRAPHY

Edwards, J. "Computer-assisted testing." Journal of Research and Development in Education, 1980, v.14, n.1.

Forster, F. "The optimum size for Rasch field testing." North west Evaluation Association Journal, May, 1977, v.1, n.2.

Forster, F. & Doherty, V. "Using the Rasch approach to measurement to solve practical school testing problems." Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Green, B. "Comments on tailored testing." In W. H. Holtzman (Ed.) Computer-Assisted Instruction, Testing, and Guidance. New York: Harper and Row. 1970.

Green, B., et.al. "Evaluation plan for the computerized adaptive vocational aptitude battery." Dept. of Psychology, Johnns Hopkins University, Baltimore, MD. May, 1982.

Lord, F. & Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lord, F. "Tailored testing: An application of stochastic approximation." Journal of the American Statistical Association. v.66, n.336, Dec., 1971.

McBride, J. "Adaptive mental testing: the state of the art." Army Research Institute for Behavorial and Social Sciences, Alexandria, VA., Nov., 1979.

McKinley, R. & Reckase, M. "A successful application of latent trait theory to tailored testing." (Research Report 80-1). Educational Psychology Dept., Univ. of Missouri, Columbia, MO, Feb., 1980.

Millman, J. "Individualizing test construction and administration by computer." In R. A. Berk (Ed.). A guide to Criterion-Referenced Test Construction. The Johns Hopkins University Press, Baltimore, MD, 1984.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut, Nielsen & Lydiche, 1960.

Rasch, G. "An item analysis which takes individual differences into account." British Journal of Mathematicl and Statistical Psychology. v.19, n.1, 1966.

Reckase, M. "Tailored testing, measurement problems, and latent trait theory." Paper presented at the annual meeting of the National Council on Measurement in Education. April, 1981.

Sampson, J. "Computer-assisted testing and assessment: current status and implications for the future." Measurement and Evaluation in Guidance. v.15, n.3, Jan., 1983.

Skinner, B. F. The Technology of Teaching. New York.: Wiley & Sons, 1968.

Spineti, J. & Hambleton, R. "A computer simulation study of tailored testing strategies for objective-based instructional programs." Educational and Psychological Measurement, 37, 1977.

Sympson, J. & Weiss, D. "Predictive validity of conventional and adaptive tests in an Air Force training environment." Dept. of Psychology, Univ. of Minnesota, Minneapolis, MN. Aug., 1982.

Waters, B. "An empirical investigation of the stratified adaptive computerized testing model." Applied Psychological Measurement. v.1, n.1, 1977.

Weiss, D. "Improving measurement quality and efficiency with adaptive testing." Applied Psychological Measurement. v.6, n.5, Fall, 1982.

Wright, B. "Solving measurement problems with the Rasch model." Journal of Educational Measurement. 14, Summer, 1977.

Wright, B., Mead, R., & Bell, S. "BICAL: Calibrating items with the Rasch model." (Statistical Laboratory Research Dept. of Education, June 1979.

Wright, B.& Stone, M. Best Test Design. Chicago.: Mesa Press. 1979.